CS 361

Probability and Statistics for Computer Science

mathematics of randomness

collection, summary & analysis of data

David Varodayan
dpv@Illinois.edu

A game of chance (win prizes!)

go.illinois.edu/cs361

About CS 361

courses.engr.illinois.edu/cs361

Datasets

A dataset $\{\boldsymbol{x}\}$ is a set of N items $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$, each of which is a tuple

e.g. our dataset consists of 78 4-tuples

Types of data

- Categorical: takes on a small set of prescribed values

  e.g. Yes/No

- Ordinal: categorical data where items can be put in order

  e.g. number of die rolls until 6 appears

- Continuous: takes on any numerical value within a range

  e.g. how much you bet

Visualizing data

- Tables
- Bar charts
- Histograms
- Conditional histograms

Summarizing data (for continuous data)

- Location parameters
  - Mean
  - Median
  - Mode
- Scale parameters
  - Standard deviation and variance
  - Interquartile range

Mean

$$\text{mean}\left(\{x_i\}\right) = \frac{1}{N}\sum_{i=1}^{N} x_i$$

Properties of the mean

- Scaling data scales the mean $\text{mean}(\{kx_i\}) = k \cdot \text{mean}(\{x_i\})$
- Translating data translates the mean $\text{mean}(\{x_i + c\}) = \text{mean}(\{x_i\}) + c$
- The sum of signed distances from the mean is 0 $\sum_i (x_i - \text{mean}(\{x_i\})) = 0$
- The mean minimizes the sum of squared distances from the data $\arg\min_\mu \sum_i (x_i - \mu)^2 = \text{mean}(\{x_i\})$

Standard deviation

$$\text{std}(\{x_i\}) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(x_i - \text{mean}(\{x_i\})\right)^2}$$

(scale)

$$= \sqrt{\text{mean}\left(\left\{\left(x_i - \text{mean}(\{x_i\})\right)^2\right\}\right)}$$

Properties of standard deviation

- Scaling data scales the standard deviation

$$\mathrm{std}(\{kx_i\}) = |k|\,\mathrm{std}(\{x_i\})$$

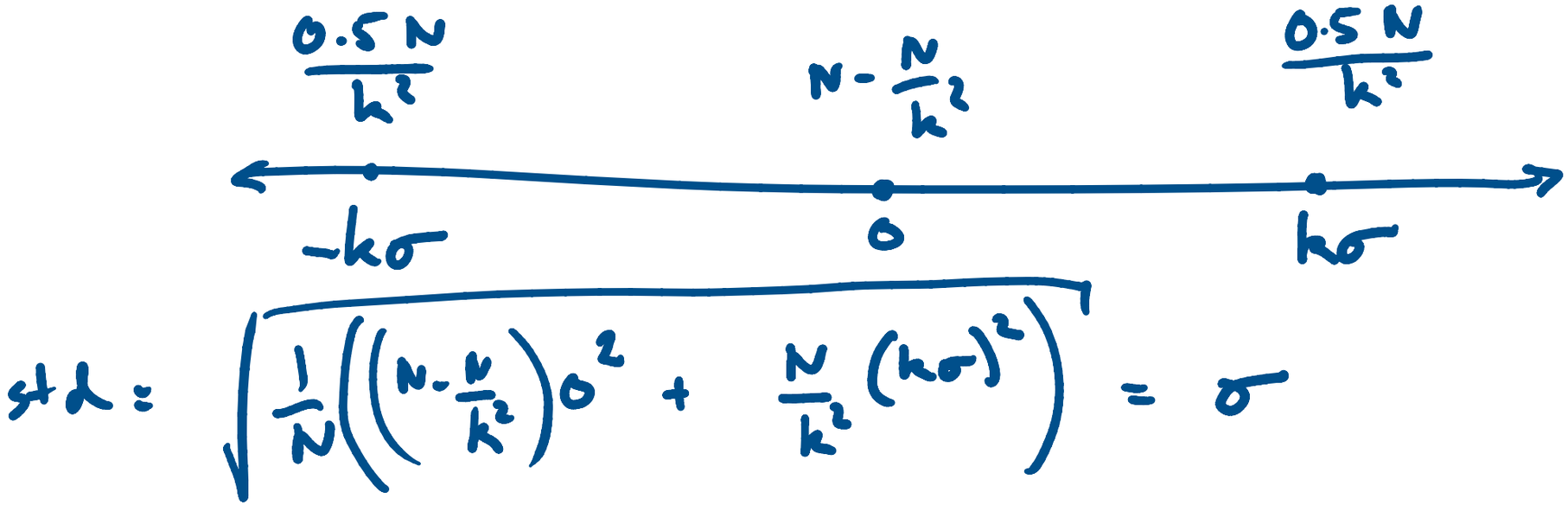
- Translating the data does **not** change the standard deviation

$$\mathrm{std}(\{x_i + c\}) = \mathrm{std}(\{x_i\})$$

Standard deviation: Chebyshev's inequality

- At most $\frac{N}{k^2}$ items are k standard deviations from the mean

- Proof: Assume the data have zero mean …



Variance = (standard deviation)^2

$$\mathrm{var}(\{x_i\}) = \frac{1}{N}\sum_i \left(x_i - \mathrm{mean}(\{x_i\})\right)^2$$

- Scaling the data scales the variance

$$\mathrm{var}(\{kx_i\}) = k^2 \cdot \mathrm{var}(\{x_i\})$$

- Translating the data does **not** effect the variance

$$\mathrm{var}(\{x_i + c\}) = \mathrm{var}(\{x_i\})$$